Reference Points and Learning*

Alan Beggs
Department of Economics and Wadham College
Oxford University
Oxford
OX1 3PN
UK
alan.beggs@economics.ox.ac.uk

December 2021

Abstract

This paper studies learning when agents evaluate outcomes in comparison to reference points, which may be adjusted in light of experience. It shows that certain models of reinforcement learning, motivated by those popular in machine learning and neuroscience, lead to classes of recursive preferences.

Keywords: reference points, reinforcement learning, recursive preferences

 and is forthcoming in the Journal of Mathematical Economics(https://doi.org/10.1016/j.jmateco.2021.102621). I am grateful to the participants in the Transatlantic Theory Workshop in Paris in September 2014, the Royal Economic Society conference in April 2016, the Department Seminar in Oxford, the Games 2016 conference in July 2016, and European Meetings of the Econometric Society in August 2016, Agnieszka Tymula, Paola Manzini and two anonymous referees for helpful comments. An early version of the paper appeared as Department of Economics discussion paper no. 767 of the same title.

1 Introduction

The idea that agents evaluate outcomes relative to reference points has been been influential in the economics literature since the pioneering work of Kahneman and Tversky (1979). It also has support in the literature in neuroscience, as outlined in the stimulating book by Glimcher (2011). How reference points are determined remains, however, an open question. Köszegi and Rabin (2006) suggest that reference points are determined by expectations and Glimcher (2011) argues that this theory is consistent with models of learning in neuroscience. This paper explores the link with these models from a theoretical perspective.

In Köszegi and Rabin (2006)'s approach expectations, and so reference points, are determined by rational expectations. In many environments the assumption of rational expectations seems too strong and it seems more reasonable to assume that expectations and reference points are determined adaptively. In the literature in neuroscience much interest has focused on modeling learning by reinforcement learning. An agent is assumed to learn in a dynamic but stationary environment of the kind studied in dynamic programming. Simple procedures can lead the agent to learn values of policies, and indeed optimal actions, even if she is unaware of the true stochastic process governing the environment and the relationship between actions and payoffs. In these papers agents form expectations and adjust them linearly according as outcomes are above or below their expectations or reference points. This paper studies generalizations of these models where the relationship between gains and losses need not be symmetric or even linear, as suggested by Kahneman and Tversky (1979) in the context of prospect theory.

Agents are assumed here to be learning about a dynamic programming problem in a stationary environment. In each state they form expectations or reference points about future payoffs and adjust these in the light of experience. Equivalently they learn about their value functions. They adjust their expectations in the manner suggested by reinforcement learning but it is assumed that they evaluate gains and losses using loss-gain functions which need not be linear. It is shown that, under mild assumptions, agents will learn to have recursive preferences. More precisely, if preferences are intertemporally separable, then if a, s and s' denote action, current and future states respectively and R and β denote payoffs and the discount factor respectively, then instead of agents' long-run value functions, V, satisfying the usual recursive equation

$$V(s) = R(a(s), s) + E_{a,s}\left(\beta V(s')\right) ,$$

they will instead satisfy an equation with the conditional expectation operator replaced by a (conditional) generalized certainty equivalent, GC:

$$V(s) = R(a(s), s) + GC_{a,s}\left(\beta V(s')\right) .$$

GC is not in general the usual conditional expectation operator but rather a (conditional) generalized certainty equivalent of the kind found when agents have non-expected utility preferences over static risky lotteries.

Such recursive preferences have become popular in macroeconomics following the work of Epstein and Zin (1989) but are sometimes regarded as rather exotic. The current paper shows they may arise as the result of boundedly rational agents learning about their environment and so provides a behavioral justification for their use. This complements the usual justification for Epstein-Zin preferences, which is purely axiomatic.

The paper also shows that under such preferences action choice can be represented as agents seeking to maximize gains or minimize losses relative to reference points as in the behavioral literature. The reference points are, however, the result of long-run learning and so agents respond rationally to shocks given their induced preferences. It also studies convergence when preferences are not intertemporally separable.

The fact that agent use reference points in their learning is taken as given. The contribution of the paper is to study how they evolve and to show how their evolution results in agents coming to have recursive preferences.

The procedures considered in the paper are simple ones. A sophisticated agent could learn about his environment more efficiently but such agents are not the focus of this paper. The purpose of the paper is instead to show that simple procedures can enable relatively unsophisticated agents to learn about their environment and in doing so come to have recursive preferences. The models in this paper are of course idealized and real agents may behave in complex ways but if they are a reasonable approximation they lend support to the idea that agents can learn to have recursive preferences.

Kuhnen (2015) provides some experimental evidence that consumers learn differently from losses and gains. In her paper she elicits beliefs from subjects and shows that agents form overly pessimistic beliefs as a result of losses. In the approach here, agents do not directly form beliefs about outcomes, rather they form expectations about future payoffs. The nonlinearity of loss-gain functions means, however, that they may revise their expectations differently in response to losses and gains. The two approaches are therefore complementary.

The models of reinforcement learning studied here are somewhat different to those familiar in the economics literature from the work of Erev and Roth

(1998). In that literature the focus is on simple rules for a single player attempting to learn in a static environment or in a static game interacting with other players. Formal analyses of the convergence properties of these models can be found in Beggs (2005) and Hopkins and Posch (2005). The models here instead analyze a single player learning in a dynamic, but stationary, environment and draw inspiration from the models of reinforcement learning in the tradition of Sutton and Barto (2020) in machine learning, which in turn have heavily influenced neuroscience.

In the neuroscience literature Niv et al. (2012) find that a model with a piecewise-linear gain-loss function may fit neural data better than conventional models. They use the model of Mihatsch and Neuneier (2002) from the machine-learning literature on risk-sensitive reinforcement learning. The current paper extends this work to general non-linear loss-gain functions and gives it an economic interpretation. In recent independent work in the machine learning literature Shen et al. (2014) have also considered non-linear loss-gain functions but they allow for a less general class than those considered here. In addition, they give a different interpretation and do not make the link with Epstein and Zin (1989) preferences.

In the economics literature Sarver (2012) considers an intertemporal model where consumers may gain utility from anticipation if they choose a high reference point but must balance this against losses from realized outcomes below the reference point. He gives an axiomatic characterization of the resulting preferences. The model here is one of learning rather optimal anticipation and the focus is on the convergence of learning schemes rather than axiomatic characterizations. Related literature is discussed in more detail in the body of the paper.

The models of reinforcement of the kind studied in Sutton and Barto (2020) have had little application in the economics literature to date. Cho and Rubinchik (2017) study a model of contemplation and intuition in a reinforcement-learning framework. There has been some application in game theory, see for example Leslie and Collins (2005), but only with standard preferences. Charpentier et al. (2021) provide a survey of reinforcement learning and possible appplications in Economics and Finance but also consider only standard preferences. Hill et al. (2021) explore using reinforcement learning to solve general equilibrium macroeconomic models where agents have standard preferences.

The paper proceeds as follows. Section 2 outlines the background on learning from neuroscience to motivate the models studied. Section 3 outlines the basic environment, which is one of stationary dynamic programming. Agents adjust their value functions evaluating a gain loss-function. Section 4 shows that these gain-loss functions correspond to estimating a

certainty-equivalent. In the long-run reference points convergence to those whose expected gain-loss is zero but this is equivalent to being a generalized certainty equivalent. These generalized certainty equivalents usually do not coincide with the standard certainty equivalent but belong to the class introduced by Chew (1989) and are those which arise in Epstein and Zin (1989) preferences. Examples of the resulting preferences are given. These include disappointment aversion (Gul (1991)) but also less familiar ones.

Section 5 studies reinforcement learning in a dynamic environment. It shows that if agents use reinforcement learning with non-linear loss-gain functions then their preferences over policies converge under mild conditions to recursive preferences of the kind introduced by Epstein and Zin (1989).

Conceptually learning to play an optimal policy has two components (a) a method to find the value of any given policy, (b) a procedure for searching for the best policy given one can determine the value of any policy. There are a number of methods one could use to implement (b), and there is some disagreement as to which is more plausible neurologically. For a fixed policy, however, most of these reduce to the same algorithm and so result in the same preferences over policies. As a result the same optimal policy will be chosen even if a different search procedure is used. For this reason although one could leap directly to learning the optimal policy, the early sections focus on learning about a fixed policy as the resulting preferences over policies will be robust to the procedure used to find the optimal policy.

Section 6 discusses learning optimal policies by using Q-learning. Conditions for convergence are given. Section 7 shows that that optimal actions can be interpreted as maximizing gains relative to reference points.

Section 8 considers extensions. In Section 5 it is assumed that preferences are intertemporally separable. Section 8 shows that this assumption can be relaxed and local convergence results obtained for general Epstein-Zin preferences. It also discusses robustness of the results to other assumptions, including the form of the reference point and the timing of shocks. Section 9 concludes.

2 Background

This section outlines background on reinforcement learning and on neuroscience in order to motivate the models studied. It may be skipped and referred back to if readers prefer to go directly to the models.

Consider a subject who receives a signal s and a random reward R, which may depend on s. In classical conditioning, interest centers on the extent to which the subject learns to predict the reward. If the prediction by the

subject at time t on receiving signal s is W_t and R_t is the reward received at time t, then a natural learning model is

$$W_{t+1} = W_t + \alpha_t(R_t - W_t)\,. \tag{1}$$

This is essentially the Rescorla-Wagner model in psychology (see for example Dayan and Abbott (2001)). That is the subject raises his prediction if the reward is greater than the prediction and lowers it otherwise. α_t is a parameter which determines the rate of adjustment.

The Rescorla-Wagner model gives a reasonable explanation of some features of conditioning (see for example Dayan and Abbott (2001)).[1] A situation it does not fit so well is one where rewards may occur at different points in time. A signal may predict that rewards will arise in future and so its occurrence may affect the agent's expectation of reward even if there is no immediate payoff.

To model this situation suppose that signals or states follow some stationary Markov chain and that the agent is interested in his total expected discounted reward from the present onwards:

$$E(\sum_{t=0}^{\infty} \beta^t R_t)\,.$$

Assume that the reward depends on the current state but not otherwise on time. If the current state is s and $V(s)$ is his expected discounted reward then a standard argument shows that

$$V(s) = R(s) + \beta E V(s')\,, \tag{2}$$

where s' is the random state tomorrow.

Suppose that at time t the state is s_t and at time $t+1$ the state is s_{t+1}.

$$\delta_t = R(s_t) + \beta V(s_{t+1}) - V(s_t) \tag{3}$$

can be thought of as an estimate of the extent to which realized payoffs differ from expectations or more poetically as a measure of disappointment or elation.

Equation (3) suggest a learning rule. Let $V_t(s)$ be the current estimate of payoffs in state s. Then a natural learning rule for updating estimates of payoffs is

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \delta_t\,. \tag{4}$$

[1]Much of its interest derives from its explanation of phenomena involving multiple stimuli, which it assumes affect rewards additively.

V is left unchanged for states other than s_t. α_t is a parameter governing the extent of adjustment each period. This is somewhat like value-function iteration except that realized values rather than expectations are used and values are adjusted only gradually.

This model, and variants, of it have attracted much attention since the work of Schultz (1998) suggesting that patterns of dopamine activation in the brain, which are thought to represent reaction to rewards, follow a pattern similar to that suggested by (4). This suggests that the brain is forming expectations of reward in the way suggested by this equation.[2]

The model in (4) is known as temporal difference learning in the literature on machine learning (see for example Sutton and Barto (2020)). If α_t is chosen appropriately then the values of V_t converge to those satisfying (2). In the learning literature this result is of considerable interest because it implies that the true values can be learned without the probabilities governing the evolution of state being known (and without any attempt to estimate them).

The model can be modified to allow for an optimal choice of action in each state, that is an optimal policy, to be learned. One popular model is so-called Q-learning. Assume that the states still follow a Markov chain but rewards and the transition probabilities depend on the current action taken. Let $Q(a, s)$ be the payoff to taking action a in state s if the optimal (stationary) strategy is followed in future;

$$Q(a, s) = R(a, s) + \beta E(V(s')|s, a)\,. \tag{5}$$

If the policy is optimal then $V(s) = \max_a Q(a, s)$ for all s so this is equivalent to

$$Q(a, s) = R(a, s) + \beta E(\max_{a'} Q(a', s')|s, a)\,. \tag{6}$$

Q-learning takes this equation and uses an analogous procedure to (4). If action a_t is played in state s_t at time t then

$$Q_{t+1}(a_t, s_t) = Q_t(a_t, s_t) + \alpha_t(a_t, s_t)\left(\beta \max_{a'} Q_t(a', s') + R(a_t, s_t) - Q_t(a_t, s_t)\right)\,. \tag{7}$$

Values of Q for other state-action pairs are left unchanged. $\alpha_t(a, s)$ is an adjustment parameter. This rule can be shown to converge to the values corresponding to the optimal policy provided sufficient experimentation is ensured. One example would be the so-called ϵ-greedy rule: with probability $1 - \epsilon$ play the action with the highest value of $Q_t(a_t, s_t)$, with probability

[2]Caplin et al. (2010) find evidence for the role of dopamine in encoding the difference between expected and realized reward using an axiomatic approach not tied to a particular learning rule.

ϵ all action are equally likely to be played. Another possibility is to choose randomly with a logit probability function (so-called softmax) : action a_t is played with probability $\exp(\delta_t Q_t(a_t, s_t)) / \sum_a \exp(\delta_t Q_t(a, s_t))$, where δ_t tends to zero at an appropriate rate.[3]

Other methods for learning the optimal action are possible and is unclear whether Q-learning is the best model for describing learning in the brain. Other less sophisticated models of learning may be more appropriate — see for example Niv and Montague (2009) for a discussion. For simplicity the paper will assume this Q-learning is used. Most of these other methods of learning are consistent with temporal difference learning for a given policy, so should result in the same preferences over policies.

Niv et al. (2012) find some evidence that a model incorporating a kind of loss aversion or risk-sensitivity may fit data from brain scans better. Following Mihatsch and Neuneier (2002) in the reinforcement learning literature they examine a variant of (4) where over-predictions are weighted more heavily than under-predictions:

$$V_{t+1}(s_t) = V(s_t) + \alpha_t \Phi(\delta_t)\,, \tag{8}$$

where

$$\Phi(x) = \begin{cases} bx & x < 0 \\ cx & x > 0\,, \end{cases} \tag{9}$$

with $b > c > 0$. That is over-predictions cause more disappointment than an under-prediction of the same magnitude causes joy. They find this fits their data better than (4) or a version in which the learning rule remains unchanged but payoffs have an expected utility form ($U(R)$).

Niv et al. (2012) and Mihatsch and Neuneier (2002) do not offer an interpretation of the value function to which this procedure converges. They also restrict attention to piecewise linear loss-functions. This paper investigates the interpretation of this learning rule and shows that the limiting value function can be interpreted as one of the Epstein and Zin (1989) class. It shows this interpretation holds for general loss functions.

3 General Model

The framework the agent operates in is the standard one of infinite horizon stationary dynamic programming:

[3] In the economics literature inspired by Erev and Roth (1998) one of the key issues is to show that the reinforcement rules used do guarantee enough experimentation — see for example Beggs (2005).

- There is infinite number of discrete time periods, $t = 0, 1, 2, \ldots$.
- Each period she must choose one of a finite number of actions from the set $A = \{1, \ldots, n\}$.
- There is a finite number of states, $S = \{1, \ldots, m\}$.
- The payoff to action a in state s is $R(a, s)$.
- If action a is chosen in state s the state will be s' next period with probability $p^a_{ss'}$
- The agent has discount factor β.

A (deterministic) stationary policy, π is a function $\pi : S \longrightarrow A$. It will also be assumed that

- Under each policy π, the set of states forms an ergodic Markov chain.

This ensures that all states are eventually visited, so it is possible to learn about them. The assumption that the number of states is finite can be relaxed by using function approximation (see for example Sutton and Barto (2020)) but will be maintained in this paper.

Conceptually learning the optimal policy consists of two parts (a) for any policy under consideration find its overall payoff, (b) search among policies to find the one with the best overall payoff. This section and the next few on issue (a): they considers a fixed policy and shows that the learning rules adopted imply that the agent will learn to value policies according to Epstein-Zin preferences. Section 6 turns to issue (b). The reader could proceed directly to Section 6 but the separation is useful as there are number of possible algorithms for searching for the optimal policy but is unclear which is more plausible neurologically. In the case of any single policy, however, they reduce to the algorithm considered here and so will result in the same preferences over policies.

Under standard preferences, the expected discounted payoff, or value, of a policy π in state s, $V^\pi(s)$, satisfies the recursive equations:

$$V^\pi(s) = R(\pi(s), s) + \beta \sum_{s'=1}^{m} p^{\pi(s)}_{ss'} V^\pi(s') \quad i = 1, \ldots, m\,, \tag{10}$$

or equivalently

$$V^\pi(s) = R(\pi(s), s) + \beta E\left(V^\pi(s')|s\right)\,. \tag{11}$$

The value function, V, of the optimal policy satisfies the Bellman equation:

$$V(s) = \max_a R(a, s) + \beta E\left(V(s')|s, a\right) . \tag{12}$$

To apply these equations to find the optimal decision rule, the agent needs to know both the transition probabilities, $p^a_{ss'}$, and the reward function R. The literature on reinforcement learning shows that the optimal rule can be learned without these being known.

Temporal-difference learning proceeds by iteratively updating an estimate of a value of an current policy π. Let V^π_t be the current estimate of V^π. Let s_t be the state at time t, a_t the action specified by policy π and s_{t+1} the state at time $t+1$. Then

$$V^\pi_{t+1}(s) = \begin{cases} V^\pi_t(s_{t+1}) + \alpha_t\left(R(a_t, s_t) + \beta V^\pi_t(s_{t+1}) - V^\pi_t(s_t)\right) & s = s_t \\ V^\pi_t(s) & \text{otherwise,} \end{cases} \tag{13}$$

where α_t is a parameter. It is assumed that both the realized reward and subsequent state are observed by the agent.

In this rule, values of estimates are only updated in the state that actually occurs (hence the second line of (13)). The advantage of this is that the learning rule is fully non-parametric but learning may be slow if the number of states is large. If one were to assume that the value function belonged to a parametric class then the experience of some states could be used to update values in other states (see for example Sutton and Barto (2020)). This would speed up learning but risks inaccuracy if the parametric assumptions are incorrect.

If $V^\pi_t = V^\pi$, then the expected change in V^π_t is zero from (12). If α_t tends to zero at an appropriate rate it can be shown that V^π_t converges to V^π, as will be discussed further in Section 4.

As discussed in the previous section, the optimal policy can be learned if temporal-difference learning is combined with an appropriate learning rule, for example Q-learning. These learning rules, however, normally reduce to temporal difference learning when considering a fixed policy. For the moment, therefore, attention is focused on how the evaluation of a given policy is affected when temporal difference learning is adapted to the case when the agent may have non-standard preferences. Search for the optimal policy will be considered in Section 6.

The term $R(a_t, s_t) + \beta V^\pi_t(s_{t+1}) - V^\pi_t(s_t)$ can be thought of as estimate of extent to which a revised estimate of future payoffs at time $t+1$ differs from the current estimate or in other words of losses and gains in comparison with expectations. Equivalently here, the agent compares current expectations

about future payoffs, $V_t^\pi(s_t) - R(a_t, s_t)$, with a revised estimate when new information is available, $\beta V_t^\pi(s_{t+1})$. These expectations are revised until the expected losses and gains are zero. A loss averse agent may weight losses and gains differently so a natural generalization would be to replace (13) by

$$V_{t+1}^\pi(s) = \begin{cases} V_t^\pi(s_{t+1}) + \alpha_t \Phi\big(R\left(a_t, s_t\right) + \beta V_t^\pi\left(s_{t+1}\right) - V_t^\pi(s_t)\big) & s = s_t \\ V_t^\pi(s) & \text{otherwise,} \end{cases} \tag{14}$$

where Φ is a function measuring losses and gains or more generally

$$V_{t+1}^\pi(s) = \begin{cases} V_t^\pi(s_{t+1}) + \alpha_t \Psi\big(\beta V_t^\pi\left(s_{t+1}\right), V_t^\pi(s_t) - R\left(a_t, s_t\right)\big) & s = s_t \\ V_t^\pi(s) & \text{otherwise,} \end{cases} \tag{15}$$

where Ψ is again a gain-loss function. In this formulation the gain (or loss) experienced if the outcome is x and the anticipated payoff, or reference point, is μ is $\Psi(x, \mu)$. Expectations of future payoffs at time t, $V_t^\pi(s_t) - R\left(a_t, s_t\right)$, are compared with a revised estimate at time $t+1$, $\beta V_t^\pi\left(s_{t+1}\right)$. That is the reference point is taken to be what is believed about future payoffs at time t. Losses and gains are caused by revisions in these beliefs. The interpretation of (15) is discussed in more detail in Sections 5 and 7.

It will be assumed that

Assumption 1. *(i)* $\Psi(x, x) = 0$ *for all* x*, (ii)* $\Psi(x, \mu)$ *is increasing in* x *and decreasing in* μ*, (iii)* Ψ *is Lipschitz in* x*, and (iv) there exist* k *and* K*,* $K \geq k > 0$*, such that for all* μ*,* μ'*,* $\mu' \neq \mu$*, and for all* x

$$k \leq \left| \frac{\Psi(x, \mu) - \Psi(x, \mu')}{\mu - \mu'} \right| \leq K\,. \tag{16}$$

In the case when $\Psi(x, \mu) = \Phi(x - \mu)$, this is implied by

Assumption 2. Φ *satisfies* $\Phi(0) = 0$ *and there exist* $m > 0$ *and* $M > 0$ *such that for all* $x \neq y$*,*

$$m \leq \frac{\Phi(y) - \Phi(x)}{y - x} \leq M\,.$$

That is Φ is Lipschitz-continuous and has slope bounded away from zero, which is clearly satisfied by the piecewise-linear form.

Recently Shen et al. (2014) have independently studied the case of gain-loss functions of the form $\Phi(x - \mu)$ and noted, as here, that Mihatsch and

Neuneier (2002)'s convergence results can be extended to them. They do not, however, study general gain-loss functions, as is done here. They offer an interpretation in terms of risk-sensitive programming with risk-measures (see for example Shapiro et al. (2014) Chapter 6 and Rusczyński (2010)) rather than the economic one given here. In particular, they do not make the link with Epstein and Zin (1989) preferences.

4 Examples of Gain-Loss Functions and Certainty Equivalents

This section gives some intuition for the convergence results in the next section. It gives examples of possible gain-loss functions and their properties and shows that the agent can be thought of as estimating a generalized certainty equivalent of the kind found in Epstein and Zin (1989) preferences. As well as expected utility, the certainty equivalents include those generated by disappointment aversion. The section can be omitted and returned to later if the reader wishes to go directly to the learning results.

The learning model (15) in the previous section has the form

$$V_{t+1}^{\pi}(s) = \begin{cases} V_t^{\pi}(s_{t+1}) + \alpha_t \Psi\big(\beta V_t^{\pi}\left(s_{t+1}\right), V_t^{\pi}(s_t) - R\left(a_t, s_t\right)\big)\big) & s = s_t \\ V_t^{\pi}(s) & \text{otherwise.} \end{cases} \tag{17}$$

That is the current reference point or expectations of future payoffs at time t, $V_t^{\pi}(s_t) - R\left(a_t, s_t\right)$, are compared with a revised estimate at time $t+1$, $\beta V_t^{\pi}\left(s_{t+1}\right)$. The value function in state s_t is revised upwards if there is a gain and downwards if there is a loss.

As well be seen in the next section, the equilibrium value function has the property that the expected gain-loss is zero in any state. That is, dropping time superscripts, if the current state is s and the state next period s' and the agent takes action $a(s)$ in state s then the equilibrium value function $V^{\pi}(s)$ satisfies

$$E_{s'} \Psi\big(\beta V^{\pi}\left(s'\right), V^{\pi}(s) - R\left(a(s), s\right)\big) = 0\,. \tag{18}$$

Suppose there are no actions to be taken but the agent receives a payoff of $\Pi(s')$ in state s'. where the distribution of s' depends on s. A gain-loss function defines a (conditional) generalized certainty equivalent CE_s by the equation

$$E_{s'} \Psi(\Pi(s'), CE_s) = 0\,. \tag{19}$$

In the current application $\Pi(s') = V^{\pi}(s')$ and CE_s is the reference point: $CE_s = V^{\pi}(s) - R(a(s), s)$. That is one can think of the reference point as a certainty equivalent. Certainty equivalents defined in this way by a gain-loss function are precisely those which correspond to Chew-Dekel preferences (Chew (1989) and Dekel (1986)) under uncertainty, which are the most general class of preferences considered in Epstein and Zin (1989). Hence the connection of reinforcement learning to Epstein and Zin (1989): the procedure considered here leads agents to have reference points and so certainty equivalents corresponding to those found in that model.

As Epstein and Zin (1989) note it is possible to consider recursive preferences with certainty-equivalents outside the Chew-Dekel class. These could not be the result of reinforcement learning. For example, Epstein and Zin (1990) consider recursive preferences with Yaari (1987)'s version of rank-dependent preferences. This version of Epstein-Zin preferences could not be the result of the learning procedure studied here.

The reminder of this section explores examples of such generalized certainty equivalents. For this purpose it is convenient to consider a fixed current state and so drop the subscript on CE and the prime on the future state, s', labeling it s instead. A generalized certainty equivalent is then defined by the equation

$$E_s \Psi(\Pi(s), CE) = 0\,. \tag{20}$$

In other words the certainty equivalent CE has the property that the expected gain-loss is zero.

Example 1. *Certainty Equivalent derived from Expected Utility*

If U is a utility function and

$$\Psi(x, \mu) = U(x) - U(\mu)$$

then CE is simply the usual certainty-equivalent:

$$CE = U^{-1}\left(E_s U\left(\Pi(s)\right)\right)\,.$$

Let GC denoted the generalized certainty equivalent defined by (20). It depends on both the reward function, $\Pi(s)$ and the probability distribution over states. In the discussion below, the latter will be held fixed, so it will be suppressed from the notation, and the solution to (20) written as $GC(\Pi(s))$. For future reference some of its properties are recorded below. The lemma below follows immediately from Assumption 1. that

Lemma 1. *(i) (constancy): if* $\Pi(s) = k$ *for all* s*, then* $GC\left(R\left(s\right)\right) = k$*. (ii) (monotonicity) if* $\Pi('s) \geq \Pi(s)$ *for all* s *then* $GC\left(R'\left(s\right)\right) \geq GC\left(R\left(s\right)\right)$*.*

The certainty equivalent will be said to be *translation-subinvariant* if $GC\left(R\left(s\right)+k\right) \leq GC\left(R\left(s\right)\right)+k$. The lemma below summarizes the well-know properties of the CCE (see for example Appendix C of Marinacci and Montrucchio (2010)).

Lemma 2. *If the certainty-equivalent is derived from expected utility then it is translation subinvariant for all risks if and only if the utility function displays increasing-absolute risk aversion.*

Some examples of certainty equivalents not derived from expected utility are:

Example 2. *Kahnemann-Tversky Form:* $\Psi(x,\mu) = \Phi(x-\mu)$*.*

In this case CE satisfies

$$E_s\Phi(\Pi(s) - CE) = 0\,. \tag{21}$$

Given that $\Phi(0) = 0$, CE represents the amount of money to be subtracted from ex post payoffs to make the decision-maker indifferent between taking and not taking the gamble. In the literature on insurance this is sometimes referred to as the 'zero-utility principle' (see for example Buhlmann (1970)). The usual certainty equivalent is the amount of money which gives the agent as much utility as taking the lottery, so is in a sense the equivalent variation, while (21) is the compensating variation.[4] Pratt (1964) refers to the certainty equivalent as the selling price of a lottery and (20) as the buying price. In general the two notions differ unless the utility function is exponential, that is displays constant absolute risk aversion.[5]

As a piece of shorthand, let CCE (compensating certainty equivalent) denote the certainty equivalent defined by (21). One can write $CE = CCE(\Pi(s))$

For future reference, some immediate properties of CCE are recorded in the following lemma.

Lemma 3. *(i) (constancy): if* $\Pi(s) = k$ *for all* s*, then* $CCE(\Pi(s)) = k$*. (ii) (monotonicity) if* $\Pi('s) \geq \Pi(s)$ *for all* s *then* $CCE\left((\Pi('s)\right) \geq CCE((\Pi(s))$*. (iii) (translation invariance) If* $\Pi('s) = \Pi(s)+k$ *for all* s *then* $CCE(\Pi('s)) = CCE\left((\Pi(s)\right)+k$*.*

[4]See for example Pope and Chavas (1985) or Kimball (1990) for further discussion.

[5]As noted for example by LaValle (1968) and Raiffa (1968).

Ranking outcomes by their compensated certainty equivalent, except in the case of constant absolute risk aversion, yields non-expected utility preferences, as will be seen in the following examples.

Example 3. *Piecewise-Linear Kahneman-Tversky Form*

Suppose Φ has the following piecewise-linear form with $0 < \lambda < 1$:

$$\Phi(x) = \begin{cases} \lambda x & x > 0 \\ x & x \leq 0\,, \end{cases} \tag{22}$$

CE is then the solution to

$$E_s I(\Pi(s) < CE)(R - CE) + \lambda E_s I(\Pi(s) > CE)(R - CE) = 0\,, \tag{23}$$

where $I(A)$ denotes the indicator function for the event A. That is CE is an expectile. An expectile can be thought of as a generalization of the idea of a quantile but applied to expectations rather than probability. It appears in the literature on estimation with asymmetric least-squares (see Newey and Powell (1987)). In particular CE minimizes the asymmetric least-squares error, where errors are weighted differently according as outcomes are less or greater than the prediction:

$$\min_{CE} E_s I(\Pi(s) < CE)(R - CE)^2 + \lambda E_s I(\Pi(s) - CE)(R - CE)^2\,. \tag{24}$$

This is consistent with the idea of loss-aversion: the decision-maker is more concerned with losses when the outcome is less than the predicted value than with gains when the outcome exceeds the prediction.

Example 4. *Disappointment Aversion*

Suppose

$$\Psi(x, \mu) = \begin{cases} U(x) - U(\mu) & x \leq \mu \\ \lambda(U(x) - U(\mu)) & x > \mu\,, \end{cases} \tag{25}$$

where U is concave and increasing and $\lambda < 1$. Then CE is the solution to

$$E_s I(\Pi(s) < CE)(U(R) - U(CE)) + \lambda E_s I(\Pi(s) > CE)(U(R) - U(CE)) = 0\,. \tag{26}$$

That is CE is the generalized-certainty equivalent in Gul (1991)'s theory of disappointment-aversion. If U is linear, $U(x) = x$, this coincides with the linear Kahneman-Tversky form. Equivalently, with a linear KT gain-loss function the two notions will coincide if outcomes, x, are measured in utils. With non-linear gain-loss functions, the two will differ.

5 Learning in a Dynamic Environment

This section and the next consider learning optimal policies. As noted in Section 3, conceptually learning the optimal policy consists of two parts (a) for any policy under consideration find its overall payoff, (b) search among policies to find the one with the best overall payoff. This section concentrates on issue (a): it considers a fixed policy and shows that the learning rules adopted imply that the agent will learn to evaluate its payoff according to a utility function belonging to the Epstein-Zin class of recursive preferences. That is the agent's preferences over policies will converge to those belonging to the of the Epstein-Zin class of recursive preferences. The next section turns to procedures to search for the optimal policy, that is step (b). One could simply consider both steps at once, and the reader should feel free to skip directly to the next section. The separation is useful conceptually, however, as there a number of algorithms that one could use to search for the best policy, and there is some debate as to which is more plausible neurologically, but they all reduce to the one in this section in the case of a fixed policy.

As laid out in section 3, the agent adopts the following learning rule:

$$V_{t+1}^{\pi}(s) = \begin{cases} V_t^{\pi}(s) + \alpha_t \Psi\left(\beta V_t^{\pi}\left(s_{t+1}\right), V_t^{\pi}(s_t) - R\left(a_t, s_t\right)\right) & s = s_t \\ V_t^{\pi}(s) & \text{otherwise.} \end{cases} \tag{27}$$

or equivalently

$$V_{t+1}^{\pi}(s) = V_t^{\pi}(s) + \alpha_t^s \Psi\left(\beta V_t^{\pi}\left(s_{t+1}\right), V_t^{\pi}(s_t) - R\left(a_t, s_t\right)\right), \tag{28}$$

where

$$\alpha_t^s = \begin{cases} \alpha_t & s = s_t \\ 0 & \text{otherwise.} \end{cases}$$

This reflects the fact that values for only one state are updated each period.

(28) may be interpreted as saying that in state s_t, $V_t^{\pi}(s_t) - R\left(a_t, s_t\right)$ are the player's expectations of future payoffs which she compares to $\beta V_t^{\pi}\left(s_{t+1}\right)$, her updated estimate of future payoffs, to determine her realized gain or loss at time $t+1$. In the special case when $\Psi(x, \mu) = \Phi(x - \mu)$, (28) can be rewritten as

$$V_{t+1}^{\pi}(s) = V_t^{\pi}(s) + \alpha_t^s \Phi\left(R\left(a_t, s_t\right) + \beta V_t^{\pi}\left(s_{t+1}\right) - V_t^{\pi}(s_t)\right). \tag{29}$$

In this case one can also interpret the player as comparing current expectations, $V^{\pi}(s_t)$, with an estimate of overall payoffs, $R(a_t, s_t) + \beta V(s_{t+1})$.

Attention will focus on the case of a fixed policy so the superscript, π, will be dropped. $a(s)$ will denote the action to be taken in state s.

It will be assumed that

Assumption 3. *$\alpha_t \geq 0$, $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.*

That is, the adjustment parameter, α_t, becomes small fast enough that the influence of random shocks dies out but not so quickly that initial conditions dominate. The assumption is satisfied by $\alpha_t = 1/t$, for example.

Standard results in stochastic approximation[6] show that the long run behavior of the system is governed by the differential equation

$$\dot{V}(s) = \delta_s E_{s'} \Psi \left(\beta V(s'), V(s) - R(a(s), s) \right) , \tag{30}$$

where δ_s reflects the amount of time spent in state s, since states are only updated one at a time. The notation $E_{s'}$ indicates that the expectation is over the distribution of states next period — the dependence of this on a and s is for convenience left implicit.

The stationary points of this equation are given by

$$E_{s'} \Psi \left(\beta V(s'), V(s) - R(a(s), s) \right) = 0 . \tag{31}$$

That is $V(s) - R(a(s), s)$ is the generalized certainty equivalent of $\beta V(s')$ given the current state s, so that in the notation of the previous section

$$V(s) = R(a(s), s) + GC_s(\beta V(s')) . \tag{32}$$

The subscript s acts as a reminder that the distribution of the state next period depends on the current state. The action taken also affects this distribution but is suppressed here as it is a function of s. β occurs inside the certainty equivalent operator but the equation can be written equivalently with it outside it.[7]

This is essentially a Bellman equation but with the certainty equivalent operator of the previous section replacing the expectation operator. It is an instance of Epstein and Zin (1989) preferences with intertemporally additive utilities. $V(s)$ can be thought of as current utility and (32) is equivalent to the usual recursive version of Epstein-Zin preferences (see equation (3.4) in Epstein and Zin (1989)) where utility is defined by a recursive equation. Here the intertemporal aggregator in Epstein and Zin (1989)'s terminology is additive so the utility function satisfies the condition that it is the sum

[6] See for example Borkar (2008) Section 7.4.

[7] Multiply the equation by β and replace βV by $\tilde{V}$ and βR by $\tilde{R}$.

of current utility plus the discounted certainty equivalent of future utility. More general intertemporal preferences are considered in Section 8.

(32) can be written equivalently as requiring that V be a fixed point of the operator T defined by:

$$TV(s) := R(a(s), s) + GC_s(\beta V(s')) \,. \tag{33}$$

T can be regarded as an operator mapping the set of bounded functions on S, $B(S)$, which is the set of all functions on S since S is finite, to itself. $B(S)$ will be given the supremum norm: $||f|| = \sup_s |f(s)|$

The following well-known result (see for example Marinacci and Montrucchio (2010)) is proven in the appendix for completeness.

Lemma 4. *Under Assumption 1 the operator T defined in* (33) *is a contraction-mapping on $B(S)$ with respect to the supremum norm if the GC is translation sub-invariant. It follows that* (32)*, which is equivalent to* (31)*, has a unique solution, $\{V^*(s)\}$, under these assumptions.*

As noted in the previous section, if the certainty-equivalent is derived from expected utility then it is translation sub-invariant if there is increasing absolute risk aversion. This is only a sufficient condition for uniqueness — see Marinacci and Montrucchio (2010) for further discussion.

Theorem 1. *Under Assumptions 1 and 3, the sequence of values V_t defined by* (27) *converge to $\{V^*\}$ with probability 1 provided the mapping T defined in* (33) *is a contraction with respect to the supremum norm. In particular the conclusion holds if the certainty equivalent is translation sub-invariant.*

The assumption of translation sub-invariance is only sufficient for the convergence and numerical results suggest it may not be necessary.

Some intuition for the result can be obtained from the fact, shown in the Appendix, that the right-hand side of (30)

$$\dot{V}(s) = \delta_s E_{s'} \Psi \left(\beta V(s'), V(s) - R(a(s), s)\right)$$

is a sign-preserving function of the right-hand side of

$$\dot{V}(s) = \delta_s \left((TV)(s) - V(s)\right) \,. \tag{34}$$

The contraction assumption implies that this latter equation is globally asymptotically stable. The argument in the Appendix shows that (30) is also globally asymptotically stable, which implies the result by standard stochastic approximation arguments.

The theorem holds in particular when the gain-loss function has the form $\Psi(x, \mu) = \Phi(x - \mu)$. In this case (32) becomes

$$V(s) = R(a(s), s) + CCE_s(\beta V(s')), \tag{35}$$

where CCE denotes the compensating certainty equivalent introduced in the last section. As noted there, CCE is translation-invariant, so Theorem 1 holds. This result was proven by Mihatsch and Neuneier (2002) in the piecewise-linear case. Although they give an interpretation in terms of dynamic programming, it is not clear from their discussion what the limiting preferences represent in economic terms. Armed with the results of the last section, one can, however, interpret the model easily.

Equations similar to (32) appear in the models of risk-sensitive control studied in the engineering literature (see for example Whittle (1996)) and introduced into economics by Hansen and Sargent (see for example Hansen and Sargent (1995) and Hansen and Sargent (2008)) but with the usual certainty equivalent. In the applications studied in engineering and by Hansen and Sargent, the certainty equivalent is usually derived from exponential utility. The formulation here allows for a general certainty equivalent and, as noted, lies in the class of Epstein and Zin (1989) preferences.

The key feature of the models in this section is that the limiting preferences can be written as satisfying equations which are linear in the probabilities (see for example (31)). These equations are therefore amenable to stochastic approximation, which essentially estimates expectations by calculating sample averages recursively, even if the probabilities are unknown. This linearity is a well known useful feature of the certainty-equivalent equations from the Chew-Dekel class. The direct recursive formulation in (32) is in general non-linear in the probabilities and so is not suitable for stochastic approximation.

As noted in the previous section, if outcomes are measured in utility then the piecewise-linear loss-gain function delivers the certainty-equivalent corresponding to Gul (1991)'s version disappointment aversion. The result in Theorem 1 implies that versions of Epstein and Zin (1989) preferences with additive utility and disappointment aversion emerge naturally as the result of simple learning processes.

6 Action Choice

The previous section considered learning the value of a given policy. If there is more than one policy with an unknown value then the agent is faced with a potentially large learning problem. There are a number of ways the

optimal policy could be learned and there is some disagreement which is more plausible neurologically. This section outlines a popular one, Q-learning, which guarantees convergence. The model of the previous section for learning about a given policy is, however, consistent with a number of procedures for selecting an optimal action and so if Q-learning is not the correct model for action-learning, preferences over policies should still have the form described previously.[8]

Recall from Section 2 that equilibrium values of $Q(a,s)$ describe the payoff to playing action a today but behaving optimally thereafter. So

$$Q(a,s) = R(a,s) + GC_{s,a}(\beta V(s'))\,. \tag{36}$$

In equilibrium $V(s') = \max_a Q(a,s')$, so this is equivalent to

$$Q(a,s) = R(a,s) + GC_{s,a}\left(\beta \max_{a'} Q(a's')\right) \tag{37}$$

or

$$E_{s'}\Psi\left(\beta \max_{a'} Q(a',s'), Q(a,s) - R(a,s)\right) = 0\,. \tag{38}$$

A version of Q-learning for this context would therefore be

$$Q_{t+1}(a,s) = Q_t(a,s) + \begin{cases} \alpha_t \Psi\left(\beta \max_{a'} Q_t(a',s'), Q_t(a,s) - R(a,s)\right) & s = s_t, a = a_t \\ 0 & \text{otherwise.} \end{cases} \tag{39}$$

In order to guarantee convergence one needs to assume there is sufficient experimentation. Let $p_t(a,s|\mathcal{F}_t)$ be the probability that action a is chosen if the system is in state s at time t, given the history of the process, $\mathcal{F}_t$ up to time t.

Assumption 4. *There exists $\eta > 0$ such that $p_t(a,s|\mathcal{F}_t) \geq \eta$ for all a, s and t and histories of the process.*

This assumption is satisfied for example by the ϵ-greedy learning (play the action with highest Q value in state s with probability $1-\epsilon$, with probability ϵ all actions are equally likely to be played) or softmax policies (a logit choice rule) described in Section 2.

As in the previous section the long-run behavior of the system is related to that of the differential equation system:

$$\dot{Q}(a,s) = \delta_{a,s}(t) E_{s'}\Psi\left(\beta \max_{a'} Q(a',s'), Q(a,s) - R(a,s)\right)\,. \tag{40}$$

[8] Other possible learning rules include SARSA and actor-critic models — see for example Sutton and Barto (2020) — but these coincide with Q-learning in the case of a fixed policy.

$\delta_{a,s}(t)$ reflects the amount of time spent in state s choosing action a, which may be time-varying as p_t may depend on time. As before, this will be globally asymptotically stable if the operator $\tilde{T}$ defined by the right-hand side of (37)

$$\tilde{T}(Q)(a,s) = R(a,s) + GC_{s,a}\left(\beta \max_{a'} R(a',s'))\right) \tag{41}$$

is a contraction on the space of bounded functions on $A \times S$. As in the previous section the contraction assumption will hold if GC is translation-subinvariant.

Theorem 2. *Under Assumptions 1, 3 and 4, Q_t converges to the unique solution of* (37)*, or equivalently of* (38)*, with probability 1 if $\tilde{T}$ is a contraction. In particular this holds if GC is translation-subinvariant.*

In particular convergence holds for loss functions of the form $\Psi(x-\mu)$, so for example for disappointment aversion.

Assumption 4 only places weak restrictions on action choice. In particular it is not necessary to play approximately the optimal policy asymptotically to learn about its value — for this reason Q-learning is often referred to as 'off-policy'. Conversely, Assumption 4, does not guarantee that the player will learn to play the optimal actions. On the other hand if the player uses a policy whose choice probabilities are continuous functions of the Q values, then convergence of the Q values implies approximate convergence of the choice probabilities to optimal levels. In particular suppose the player uses an ϵ-greedy policy:

Assumption 5. *For all t and s_t the player chooses the action a which maximizes $Q_t(a,s_t)$ with probability $1-\epsilon$ and plays a randomly chosen action with probability ϵ, all actions being equally to be chosen.*

It follows immediately from Theorem 2 that

Corollary 1. *Under Assumptions 1, 3 and 5, for any $\epsilon' > 0$, there exists $\epsilon > 0$ in Assumption 5 such that*

$$\lim_{t\to\infty} P(\textit{optimal action played at time } t \textit{ in state } s_t) \geq 1-\epsilon'$$

In other words, asymptotically the player will choose optimal actions with arbitrarily high probability if ϵ is small enough. η in Assumption 4 and ϵ in Assumption 5 can probably be allowed to decay to zero at an appropriate rate, so exactly optimal actions are played asymptotically. This is known in the standard case — see for example Singh et al. (2000) — but the proof

technique here with nonlinear loss-gain functions is different and not so easily adapted to show this. η can, however, be arbitrarily small.

This algorithm only updates the values for a state when that state is actually visited. If the number of states is large this may mean slow learning. One could instead assume that the Q functions belong to a particular parametric class so that payoffs at one state are informative about those at others (see for example Sutton and Barto (2020)). These would speed up learning, though agents may now only learn approximately optimal policies if the parametric class does not contain the true Q function.

The convergence result requires that agents experiment sufficiently Assumption 4) and that learning rates go to zero at an appropriate rate (Assumption 3). If agents cease experimenting too quickly or learning rates go to zero too quickly then agents may learn play policies which are at best locally optimal. If learning rates go to zero too slowly then random shocks will perpetually affect the agents' decisions. The model here is clearly an idealization and actual learning processes may be more complex but if it is a reasonable approximation then it lends support to the idea that agents might come to have recursive preferences and learn to play optimal or approximately optimal actions.

Whether Q-learning is the best model of learning of the brain is debated in the neuroscience literature. If it is then the results here show that the optimal policy can be learned for a wide range of recursive preferences.

7 Implications for Behavior

This section considers the implications of the model for behavior. It shows that choices can thought of maximizing gains about some reference point and compares this to other formulations of reference points.

The Bellman equation for optimal action choice is:

$$V(s) = \max_a \; R(a,s) + GC_{s,a}(\beta V(s'))\,. \tag{42}$$

This has a unique solution if GC is translation-subinvariant. The subscripts s and a on GC reflect the fact that both the state and action affect the distribution of states next period.

The Bellman equation can be rewritten in the form of an agent maximizing gains relative to a reference point. Indeed, it is equivalent to V solving the set of equations:

$$\max_a E_{s'} \Psi\left(\beta V(s'), V(s) - R(a,s)\right) = 0\,. \tag{43}$$

So in particular in each state s the optimal action $a(s)$ solves

$$a(s) \quad \max_a E_{s'} \Psi \left(\beta V(s'), V(s) - R(a,s) \right) . \tag{44}$$

That is the agent maximizes expected gains in comparison to the reference point, $V(s) - R(a,s)$. The discussion can be summarized in the following Lemma (proof in Appendix).

Lemma 5. (42) *and* (43) *have the same solutions. The solution is unique if GC is translation-subinvariant.*

In the case of preferences of the Kahneman-Tversky form (44) has a particularly pleasing form. It implies that the optimal action solves

$$a(s) \quad \max_a \; E_{s'} \; \Phi \left(R(a,s) + \beta V(s') - V(s) \right) . \tag{45}$$

This gives an intuitive representation in terms of loss-aversion. The optimal action maximizes the expected gain relative to the reference point $V(s)$. Note, however, that the agent's behavior is forward-looking in that she takes into account future losses and gains and not directly whether the current period is good or bad. This is made clear in the formulation of (42). To reconcile with (45) note that $V(s)$ already incorporates the fact that the agent may be in a disappointing state. Evidence of this kind of forward-looking loss-aversion can be found in the recent experiments of Abeler et al. (2011) and Gill and Prowse (2012).

If the agent did not adjust her reference point in response to the state then (45) would become

$$a(s) \quad \max_a \; E_{s'} \; \Phi \left(R(a,s) + \beta V(s') - \bar{V}) \right) \tag{46}$$

or even

$$a(s) \quad \max_a \; E_{s'} \; \Phi \left(R(a,s) + \beta \bar{V} - \bar{V}) \right) . \tag{47}$$

If Φ is sufficiently concave the agent might then choose to work less hard if the state were relatively good even though the marginal payoff to effort is higher — compare the literature on income targets in labor supply originated by the work of Camerer et al. (1997). In the forward-looking case this is less clear as the effect will depend on how curvature of Φ affects the marginal impact of effort, in a given state, on the certainty equivalent in future — see (42).

Köszegi and Rabin (2009) suggest a model where past expectations can affect current behavior. Their model is complex but a simplified version in the current context would be to assume that agents act as in (46) with

$\bar{V}$ equal to their expectations in the previous period ($V(s_{t-1})$). Agents will then act to avoid disappointing their past expectations. Such preferences will, however, in general lead to time-inconsistent behavior, whereas the current formulation leads to time-consistent behavior.[9]

Sarver (2012) proposes a theory of reference points resulting from optimal anticipation: consumers derive utility from the anticipation caused by a high expectation or reference point but must balance this against the loss caused by disappointment of the outcome being below the reference point. Here consumers do not gain utility from anticipation. Sarver (2012) gives an axiomatic characterization of preferences. The approach here is not axiomatic but asks what preferences may emerge if the decision-maker adjusts reference points to minimize expected losses.

Köszegi and Rabin (2006) and Köszegi and Rabin (2007) suggest that reference points are determined by expectations. They examine, however, a more complex notion where agents consider the entire distribution of returns rather than a single reference point. Gains and losses are evaluated by comparing the realized outcome to every outcome considered possible. A detailed discussion of the relationship of their model to other theories can be found in Masatlioglu and Raymond (2014).

8 Extensions

This section discusses some extensions and limitations of the results. The first sub-section shows that the results can be extended to Epstein-Zin preferences with non-additive intertemporal preferences. The second discusses the case of random payoffs and the issues of time consistency encountered if these are allowed. The third considers the form of the reference point, and implications for time consistency, and the fourth robustness to other assumptions.

8.1 Non-additive Intertemporal Preferences

The full family of Epstein-Zin preferences allows for non-additive intertemporal preferences and these are commonly used in applications. It is shown in this section that these can also be interpreted as the outcome of a learning rule.

Epstein and Zin (1989) assume that recursive preferences have the form

$$V(s) = W(R(a(s), s), GC_s(V(s'))\,, \tag{48}$$

[9]Further discussion of time-consistency can be found in Section 7.

where $W(x,y)$ is an inter-temporal aggregator and GC_s a generalized certainty equivalent. A popular choice for the aggregator W is the CES form

$$W(x,y) = \left(x^\rho + by^\rho\right)^{1/\rho}. \tag{49}$$

It will be assumed that

Assumption 6. *W is increasing in x and y and for some $\beta < 1$, $|W(x,y) - W(x,y')| \leq \beta|y - y'|$ for all x,y,y'. GC is translation-subinvariant.*

This is a standard assumption in the literature on recursive preferences (see for example Marinacci and Montrucchio (2010)) and the assumption on W is satisfied by the CES form if $\rho > 1$ and $b < 1$. It guarantees that there is a unique solution to (48).

Provided W is invertible in the second argument one can apply stochastic approximation to this form of preferences. If $z = W(x,y)$ let $y = \tilde{W}(x,z)$ be the inverse function. $\tilde{W}$ is well-defined in the CES case.

(48) can be re-written as

$$\tilde{W}(R(a(s),s),V(s)) = GC_s(V(s')) \tag{50}$$

or equivalently

$$E_{s'}\Psi\left(V(s'), \tilde{W}\left(R\left(a\left(s\right),s\right),V\left(s\right)\right)\right) = 0 \tag{51}$$

and one can apply the learning scheme

$$V^\pi_{t+1}(s) = \begin{cases} V^\pi_t(s_{t+1}) + \alpha_t \Psi\left(V^\pi_t\left(s_{t+1}\right), \tilde{W}\left(R\left(a_t,s_t\right), V^\pi_t(s_t)\right)\right) & s = s_t \\ V^\pi_t(s) & \text{otherwise.} \end{cases} \tag{52}$$

As in the previous case, the asymptotic behavior of the system is related to that of a differential equation, in this case

$$\dot{V}(s) = \delta_s E_{s'}\Psi(V(s'), \tilde{W}(R(a(s),s),V(s)), \tag{53}$$

where again δ_s denotes the long-run proportion of time spent in state s.

In principle the arguments of the previous sections can be extended to show global convergence of this procedure. $\tilde{W}$ need not, however, exist everywhere — consider for example the CES case with x much larger than z — and this makes the algorithm hard to define globally. For this reason a local result will be given.

Let subscripts denote partial derivatives.

Assumption 7. *In addition to Assumption 6, $\tilde{W}$ is well-defined in a neighborhood of any solution to* (48). *Ψ and $\tilde{W}$ are C^1, and $\tilde{W}_2 \geq \tilde{\beta} > 1$ for all x,y, some $\tilde{\beta}$.*

These are standard assumptions in the literature on recursive preferences (see for example Marinacci and Montrucchio (2010), who refer to this as the Blackwell case). W satisfies this in the CES case if $\rho > 1$ and $b < 1$. The C^1 assumption rules out kinks but such a loss function can be well approximated be a smooth one which is very curved near the kink points.

Assume also

Assumption 8. *$\alpha_t = \frac{1}{M+t}$, for some $M > 0$.*

The parameter M allows one to control the step size, to ensure that initial random shocks are unlikely to take the process out of the domain of attraction of the equilibrium.

One then has the following result.

Theorem 3. *Under Assumptions 7 and 8, for any $\epsilon > 0$, there is a neighborhood of the unique solution to* (51), *$\{V^*(s)\}$, such that if the initial values of V lie in this neighborhood and M is large enough, then the values of V_t generated by* (52) *converge to V^* with probability at least $1 - \epsilon$.*

Thus general Epstein-Zin preferences can be regarded as the outcome of a learning procedure, albeit with a more complex procedure. The results can also be extended to action-learning.

8.2 Random Payoffs and Time-Consistency

It has been assumed that rewards are a deterministic function of the current state. In dynamic programming it is common to assume that rewards may be random. For example, one could assume as in Stokey et al. (1989), write the reward as a function of the current and future state, $R(a, s, s')$. One can still apply the learning algorithm to this case, as do Mihatsch and Neuneier (2002), but the resulting preferences will be time-inconsistent in general.

In more detail

Assumption 9. *The assumptions are unchanged except current rewards are a function of the current and future state: $R(a, s, s')$.*

For simplicity consider the case of preferences of the Kahneman-Tversky form ($\Phi(x - \mu)$). The issues are the same in the general case.

Theorem 4. *Under Assumption 9, the values of V_t generated by* (14) *converge with probability 1 to the unique solution of*

$$E_{s'}\Phi\left(R(a(s),s,s')+\beta V(s')-V(s)\right)=0\,. \tag{54}$$

(54) is equivalent to

$$V(s)=CCE\left(R(a(s),s,s')+\beta V(s')\right)\,. \tag{55}$$

If the latter could be written as

$$V(s)=ER(a(s),s,s')+CCE\left(\beta V(s')\right) \tag{56}$$

then preferences would be time-consistent.

R cannot, however, be extracted from the CCE even in expectation in general. Unlike (35), therefore, (55) is in general not weakly separable between current and future states, so preferences are not recursive. Decisions will therefore be time-inconsistent in general (see Johansen and Donaldson (1985) for example). Preferences will be time-consistent provided the state can be defined in such a way that any randomness regarding payoffs is resolved in the next period rather than the current one. As is well understood in the Epstein and Zin (1989) and Kreps and Porteus (1978) framework, when uncertainty is resolved is crucial once one goes beyond additively separable expected utility preferences.

8.3 Form of Reference Point in the General Case

In the case of preferences of the Kahneman-Tversky form the loss or gain experienced by the agent if the state is s' is $\Phi(R(a,s)+\beta V(s')-V(s))$. One could interpret this in two ways. One could regard the reference point, $V(s)$, as measuring total expected payoffs in state s and so it is compared with current payoff plus a revised estimate of future payoffs, $R(a,s)+\beta V(s')$. Another interpretation would be that as payoffs in state s are known the reference point $V(s)-R(a,s)$ measures expected future payoffs and this is compared with a revised estimate once state s' is known. Either interpretation yields the same limiting values for V.

Consider the general case, with time-separable preferences for simplicity. Here the paper has implicitly taken the second interpretation: the gain is $\Psi(\beta V(s'),V(s)-R(a,s))$. If one were instead to take the first interpretation the relevant gain would be $\Psi(R(a,s)+\beta V(s'),V(s))$. In this case expected gain is zero when

$$V(s)=GC\left(R(a(s),s)+\beta V(s')\right)\,. \tag{57}$$

In this case preferences are no longer of Epstein-Zin form. As in the previous subsection preferences are in general no longer weakly separable, so will be time-inconsistent. In the Kahneman-Tversky case the certainty-equivalent is translation-invariant so

$$CCE\left(R(a(s),s)+\beta V(s')\right)=R(a(s),s)+CCE(\beta V(s'))\,. \tag{58}$$

and the two formulations are equivalent.

Which is more behaviorally appealing in general is open to debate. The view taken here is that once state s is known the agent revises expectations to take into account information on payoffs accrued, so the second is utilized.

8.4 Robustness to Other Assumptions

8.4.1 Assumptions on Preferences and Loss Functions

The assumption that the certainty-equivalent associated with Ψ is translation sub-invariant is quite strong but is fairly standard in the literature on recursive preferences. As is clear from the statement of Theorem 1, it is only a sufficient condition for convergence and can be weakened provided the Bellman operator remains a contraction.

The Lipschitz assumptions on Ψ rules out gain-loss functions with infinite slopes at the origin, say of a fractional power form, but these can be arbitrarily well approximated by gain-loss functions with large but finite slopes at the origin. They also bound the growth rate from above and below ruling out say exponential or bounded loss functions. It is enough, however, that the assumptions apply in the domain of interest. Under current assumptions, the algorithm cannot leave a certain bounded region containing the optimal payoffs (see for example the proof of Theorem 1), so the restriction need only hold here. If random shocks, as in Section 7.2, mean that the system may leave this region one could modify the algorithm by truncating it to ensure that it always remains in it, as is common in reinforcement learning — see for example Kushner and Yin (1997) Chapter 12.8 for details.

8.4.2 Step Sizes

The rule for step-sizes in (14) can be generalized. One can for example allow the update rate for each state to depend on the number of times it has been visited rather than being linked to the total number of periods that have elapsed. One could for example in (28) set

$$\alpha^s(t)=\begin{cases}1/(n_s(t)+1) & \text{if } s \text{ is visited at } t\\ 0 & \text{otherwise.}\end{cases} \tag{59}$$

where $n_s(t)$ is the number of visits to state s by time t. This is referred to in the literature as asynchronous updating (see for example Borkar (2008) Chapter 7).

Theorem 1 would remain valid with only small changes in the proof[10] provided with probability 1, for some constant C, for all s

$$\sum_t \alpha^s(t) = \infty \qquad \text{and} \qquad \sum_t (\alpha^s(t))^2 \leq C\,.$$

This requires that each state is visited infinitely often, which holds here since the chain is ergodic. It is satisfied by (59). A similar generalization can be performed for Q-learning.

9 Conclusion

This paper has studied models of learning where agents compare outcomes to reference points and adjust their reference points in light of outcomes. It has shown that such a process can lead agents to have recursive preferences and so strengthened the case for their use.

[10] In particular the constants δ_s which reflect relative spent in each state would no longer appear.

Appendix

Proof of Lemma 4

Consider the following map defined on the set of bounded functions on S, $B(S)$, with the supremum metric,

$$T(V)(s) = R(a(s), s) + GC_s(\beta V(s'))\,. \tag{60}$$

Since S is finite R is bounded (and $B(S)$ is in fact just the set of real-valued functions on S). Also by the constancy and monotonicity properties of GC_s (Lemma 1), if V is uniformly bounded then so is $GC_s(V(s'))$. Hence T maps $B(S)$ to itself.

It follows from the monotonicity property of GC_s (Lemma 1) that

$$V(s) \geq W(s)\ \forall s \Longrightarrow TV(s) \geq TW(s)\ \forall s \tag{61}$$

and from the assumption of translation-subinvariance that

$$T(V + k)(s) \leq T(V)(s) + \beta k \qquad \forall s\,. \tag{62}$$

(61) and (62) imply that T satisfies Blackwell's sufficient condition for a contraction on $B(S)$ (see for example Stokey et al. (1989) p. 54, Theorem 3.3). Hence T has a unique fixed point, which is equivalent to uniqueness of the solution to (32). The equivalence of (32) and (31) follows from the remarks in the text.

Proof of Theorem 1

The learning rule can be written as

$$V_{t+1}(s) = V_t(s) + \alpha_t I(s = s_t)\left(F_s\left(V_t\right) + u_{t+1}(s)\right)\,, \tag{63}$$

where $I()$ is an indicator function and F_s the expected change in V_t in state s (unweighted by α_t):

$$F_s(V) = \sum_{s'} p_{ss'}^{a(s)}\ \Psi\left(\beta V\left(s'\right), V\left(s\right) - R\left(a\left(s\right), s\right)\right) \tag{64}$$

and $u_{t+1}(s) = \Psi\left(\beta V^t\left(s_{t+1}\right), V(s) - R\left(a(s), s\right)\right) - F_s(V^t)$ is the random error.

If $\mathcal{F}_t$ denotes the history of the process up to and including time t, then by construction of u_t, for all t

$$E(u_{t+1}(s)|\mathcal{F}_t) = 0 \tag{65}$$

and also there exists some $M > 0$ such that for all t

$$E(u_{t+1}^2(s)|\mathcal{F}_t) \leq M\left(1 + \sup_{\tau \leq t, s} |V_\tau(s)|^2\right)\,. \tag{66}$$

The second property uses boundedness of payoffs and the fact, which will be used repeatedly, that

$$|\Psi(x,y)| \leq K|x-y|) \qquad \forall x,y\,. \tag{67}$$

This follows from Assumption 1 since $\psi(x,x)=0$ and ψ is Lipschitz in its second argument with constant K.

Since F_s is Lipschitz, by Assumption 1, and (65) and (66) hold, Theorem 2 of Chapter 7 and the remarks in the final paragraph of Section 7.4 of Borkar (2008) show that V_t converges to V^* almost surely if the system of differential equations

$$\dot{V}(s) = \delta_s F_s(V) \tag{68}$$

is globally asymptotically stable provided that V_t is almost surely bounded. δ_s denotes the proportion of time spent in state s under the current fixed policy (note that the chain is ergodic).

Proof of the boundedness of V_t will be deferred. First asymptotic stability of (68) will be shown.

Step 1: Global asymptotic stability of (68)

It will be shown that the function $\Upsilon(s) = \sup_s |V(s) - V^*(s)|$ is a non-smooth Lyapounov function for (68).

Consider a differential equation $\dot{x} = f(x)$ with equilibrium point x^*. Let $W(x)$ be a non-negative Lipschitz-continuous function with (i) $W(x^*)=0$, (ii) $m\|x-x^*\| \leq W(x) \leq M\|x-x^*\|$ for some $m, M > 0$. $\|\cdot\|$ denotes the supremum norm - note that all norms on Euclidean space are equivalent. According to Theorem 6.3 in Chapter II of Rouche et al. (1977)[11] the equilibrium point, x^*, of the differential equation $\dot{x} = f(x)$ is globally asymptotically stable if for some $\mu > 0$, for any trajectory

$$D^+W(x(t)) \leq -\mu\|(x(t)-x^*\| \qquad \forall t \geq 0\,, \tag{69}$$

where D^+ is the upper-Dini derivative along the trajectory[12]:

$$D^+W(x(t)) = \limsup_{h\downarrow 0} \frac{W(x(t)+hf(x(t)) - W(f(x(t)))}{h}\,. \tag{70}$$

Υ clearly satisfies (i) and (ii), so it remains to verify (70).

By definition $GC_s(\beta V(\tilde{s}))$ satisfies[13]

$$\sum_{s'} p_{ss'}^{a(s)}\ \Psi\left(\beta V(s'), GC_s(\beta V(\tilde{s}))\right) = 0\,.$$

[11]The result there is slightly more general but the formulation here is sufficient for current purposes.

[12]See Theorem 4.3 in Appendix I of Rouche et al. (1977).

[13]$\tilde{s}$ is used as generic representative of the state next period so that is clear that GC is independent of the actual realized state, s'.

Hence (68) is equivalent to

$$\dot{V}(s) = \delta_s \sum_{s'} p_{ss'}^{a(s)} \Big[\Psi(\beta V(s'), V(s) - R(a(s), s)) - \Psi(\beta V(s'), GC_s(\beta V(\tilde{s}))) \Big] . \tag{71}$$

Since Ψ is Lipschitz in its second argument, by the Intermediate Value Theorem for Lipschitz functions (see Clarke (1990) Theorem 2.3.7) there are $k_{ss;'V}$ dependent[14] on s, s' and V in $[k, K]$ (see Assumption 1 for the definition of $k > 0$ and K) such that

$$\dot{V}(s) = \delta_s \sum_{s'} p_{ss'}^{a(s)} k_{ss';V} \big(R(a(s), s) + GC_s(\beta V(\tilde{s})) - V(s) \big) \tag{72}$$

or equivalently

$$\dot{V}(s) = \delta_s \left(\sum_{s'} p_{ss'}^{a(s)} k_{ss';V} \right) ((TV)(s) - V(s)) , \tag{73}$$

where T is the contraction operator introduced in Section 5.

It is straightforward to see that

$$D^+(\Upsilon) = \max_{s \in I(V)} \operatorname{sgn}(V(s) - V^*(s)) \, \delta_s F_s(V) , \tag{74}$$

where $I(V)$ is the set of indices s for which $|V(s) - V^*(s)|$ is greatest.

Let the modulus of contraction of T be $\lambda < 1$. Since $(TV^*)(s) = V^*(s)$ it follows that if $s \in I(V)$ then $\operatorname{sgn}(V(s) - V^*(s)) (TV(s) - V(s)) \leq (\lambda - 1))|V(s) - V^*(s)|$. Using (73) it follows that

$$D^+(\Upsilon) \leq -\delta k (1 - \lambda) \sup_s |V(s) - V^*(s)| , \tag{75}$$

where $\delta = \inf_s \delta_s$ and k is the constant in Assumption 1.

It follows that Υ satisfies (69) with $\mu = \delta k(1 - \lambda)$ and hence (68) is globally asymptotically stable.

Step 2: Boundedness of V_t.

From Assumption 3, $\alpha_t \to 0$ as $t \to \infty$. Let τ be such that $\alpha_t K < 1$ for all $t \geq \tau$ where K is the constant in Assumption 1.

Let $M = \max\{\max_s V_\tau(s), \max_s R(a(s), s)\}$. It will be shown that

$$V_t(s) \leq \frac{M}{1 - \beta} \qquad \text{for all } s \text{ and } t \geq \tau , \tag{76}$$

[14]Dependence on R and $a(s)$ is not indicated explicitly as these are both functions of s.

which establishes boundedness. The argument is a simple adaptation of that of Gosavi (2006) in the standard case.

The result is true by definition when $t = \tau$. The result will be established by induction.

For any t and s

$$V_{t+1}(s) = \begin{cases} V_t(s) + \alpha_t \Psi \left(\beta V_t(s_{t+1}), V_t(s) - R(a(s), s) \right) & s = s_t \\ V_t(s) & \text{otherwise.} \end{cases} \tag{77}$$

In the second case, if (76) holds for V_t then $|V_{t+1}(s)| \leq M/(1-\beta)$ trivially. In the first case, when $s = s_t$, it follows, as earlier in the proof, from Assumption 1 and the intermediate value theorem for Lipschitz functions that

$$\begin{aligned} & V_{t+1}(s_t) \\ & = V_t(s_t) + \alpha_t \left(\Psi \left(\beta V_t(s_{t+1}), V_t(s_t) - R(a(s_t), s_t) \right) - \Psi(\beta V_t(s_{t+1}), \beta V_t(s_{t+1})) \right) \\ & = V_t(s_t) + \alpha_t \kappa \left(\beta V_t(s_{t+1}) - V_t(s_t) + R(a(s_t), s_t) \right) \\ & = (1 - \alpha_t \kappa) V_t(s_t) + \alpha_t \kappa \beta V_t(s_{t+1}) + \alpha_t \kappa R(a(s_t), s_t) \end{aligned}$$

for some $\kappa \in [k, K]$. Since $t \geq \tau$, $\tilde{\alpha}_t = \alpha_t \kappa < 1$. Using the inductive hypothesis and the definition of M it follows that

$$V_{t+1}(s_t) \leq (1 - \tilde{\alpha}_t) \frac{M}{1-\beta} + \beta \tilde{\alpha}_t \frac{M}{1-\beta} + \tilde{\alpha}_t M = \frac{M}{1-\beta} .$$

It follows, combining the two cases in (77), that if (76) holds for t it holds for $t+1$, so the inductive step is complete.

Combining Steps 1 and 2 proves the Theorem.

Proof of Theorem 2

The proof is almost exactly the same as that of Theorem 1 except that the relevant differential equation is now

$$\dot{Q}(a, s) = \delta_{a,s}(t) E_{s'} \Psi \left(\max_{a'} Q(a', s'), Q(a, s) - R(a, s) \right) . \tag{78}$$

As in Step 1 above, (78) can rewritten in the form

$$\dot{Q}(a, s) = \delta_{a,s}(t) \sum_{s'} p^a_{ss'} k_{s,s',Q} \left((\tilde{T}Q)(a, s) - Q(a, s) \right) . \tag{79}$$

Assumption 4 and the fact that the states form an ergodic chain under any policy imply that $\delta_{a,s}(t)$ is uniformly bounded below by some $\underline{\delta}$ (see for example Borkar (2008) Section 10.3 or, for a more detailed proof, Lemma

A1 of Perkins and Leslie (2012)). Using this, if $\tilde{T}$ is a contraction, global asymptotic stability follows using the Lyapounov function $\max_{s,a} |Q(s,a) - Q^*(s,a)|$ as in the proof of Theorem 1.

Boundedness is established in a similar way to Step 2 of Theorem 1.

The fact that $\tilde{T}$ is a contraction if the certainty equivalent is translation sub-invariant follows from the proof of Lemma 5.

Proof of Lemma 5

The proof of uniqueness follows from the fact that the operator:

$$\tilde{T}(V)(s) = \max_a R(a,s) + GC_{a,s}(\beta V(s')) \tag{80}$$

is a contraction on the set of bounded functions on $A \times S$, $B(A,S)$, endowed with the supremum metric if GC is a translation-subinvariant. The proof is almost identical to that of Lemma 4.

The equivalence follows since for any real number k and random variable X:

$$k \leq GC(X) \iff E\Psi(X,k) \geq 0\,,$$

so $V(s) - R(a,s) \geq GC_{s,a}(\beta V(s'))$ is equivalent to $E_{s'}\Psi(\beta V(s'), V(s) - R(a,s)) \leq 0$.

Proof of Theorem 3

The uniqueness of the solution to (51) follows from a similar argument to that in Lemma 4. The assumed properties of W and GC imply that the operator $\hat{T}$ defined by

$$\hat{T}(V)(s) = W\left(R(a(s),s), GC(V(s'))\right) \tag{81}$$

maps bounded functions to bounded functions and satisfies Blackwell's sufficient conditions for a contraction.

The current algorithm can be written in the form

$$V_{t+1} = V_t + \alpha_t H(V, \xi_t)\,, \tag{82}$$

where $\xi_t = (s_t, s_{t+1})$ and H is the vector-valued function obtained from the right-hand side of (52):

$$H(V,(s,s'))(\tilde{s}) = \begin{cases} \Psi(V(s'), \tilde{W}(R(a(s),s), V(s))) & s = \tilde{s} \\ 0 & \text{otherwise.} \end{cases}$$

s_t is an ergodic Markov chain with finitely many states. Let $h(V)$ be the expectation of H with respect to the stationary distribution, δ_s of s_t, that is

$$h(V) = \delta_s \sum_{s'} p_{ss'}^{a(s)}\ \Psi(V(s'), \tilde{W}(R(a(s),s), V(s)))\,. \tag{83}$$

Let Q be a compact subset of the domain of attraction of a locally asymptotically stable equilibrium, V^*, of

$$\dot{V} = h(V)\,. \tag{84}$$

(82) is a special case of the algorithms studied by Benveniste et al. (1990) and Theorem 13 of Chapter 1 of Part II of their book shows that there is a constant $B(Q)$ such that

$$\Pr(V_t \to V^* | V_0 \in Q) \geq 1 - B(Q) \sum \alpha_t^2 \tag{85}$$

provided their assumptions (A1) to (A7) are met. Since Assumption 8 implies that $\sum_t \alpha_t^2$ can be made arbitrarily small, to prove the result it is enough to verify these assumptions and show that V^* is locally asymptotically stable under (84).

Verification of (A1) to (A7) is straightforward. (A1) and (A6) require that α_t be decreasing and $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$ and so are implied by Assumption 8. (A2) requires that ξ_t be Markov for each fixed value of $V = V_t$, which follows since s_t is Markov and hence so is ξ_t (and the transition laws are independent of V). (A3) requires that $H(V, \xi)$ be uniformly bounded by a polynomial in ξ when V lies in any compact set, which is immediate from Assumption 7 and the finiteness of the chain. Similarly (A5) imposes growth conditions on the conditional moments of ξ_{t+1}, which are immediate here as the chain is finite and independent of V. (A4)(i) requires that h is Lipschitz, which follows from Assumption 7, and that the so-called Poisson equation $(I - \Pi_V)\nu_V(\xi) = H(V, \xi) - h(V)$ have a unique solution ν_V which is Lipschitz in V. This follows from Ma et al. (1990) Theorem 4.3 since ξ_t is an ergodic Markov chain with transition laws independent of V (so in particular ξ_t is aperiodic with a single recurrent class for each V and its transition probabilities are Lipschitz, which are their assumptions (C1) and (C2), and H is Lipschitz from Assumption 7, which is their assumption (C3)).

(85) will therefore imply the result provided it can be shown that V^* is locally asymptotically stable. (84) written out in full is

$$\dot{V}(s) = \delta_s \sum_{s'} p_{ss'}^{a(s)}\ \Psi(V(s'), \tilde{W}(R(a(s), s), V(s)))\,. \tag{86}$$

The Jacobian matrix of the right-hand side, $G(V)$, is a dominant-diagonal matrix at V^*. To prove this note that $\Psi_1 \geq 0$ and $\Psi_2 < 0$ (subscripts denoting derivatives). Moreover if μ is the generalized certainty equivalent of a distribution $x_1, \ldots, x_n$ with probabilities $q_1, \ldots, q_n$ then if Ψ is sub-invariant

$$\sum_\sigma q_\sigma \left(\Psi_1(x_\sigma, \mu) + \Psi_2(x_\sigma, \mu)\right) \leq 0\,. \tag{87}$$

Now for each s, $\tilde{W}(R(a(s),s),V^*(s))$ is the generalized certainty equivalent of $\{V^*(s')\}$ conditional on s. Applying (87) to the right-hand side of (86) yields, noting that $\tilde{W}_2 > 1$, that for any s at V^*

$$-\sum_{s'} p_{ss'}^{a(s)} \Psi_2|_{s'} \tilde{W}_2 - p_{ss}^{a(s)} \Psi_1|_s > \sum_{s' \neq s} p_{ss'}^{a(s)} \Psi_1|_{s'} \,, \tag{88}$$

where, with some abuse of notation, $\Psi_1|s'$, for example denotes $\Psi_1(V^*(s'), \tilde{W}(R(a(s),s),V^*(s)))$, similarly for $\Psi_2|_{s'}$ and $\tilde{W}_2$ denotes $\tilde{W}_2(R(a(s),s),V^*(s))$. That is, the Jacobian matrix G is diagonally-dominant with a negative diagonal at V^*.

Now the eigenvalues of a dominant-diagonal matrix with strictly-negative diagonal have negative real parts. V^* is therefore locally asymptotically stable.

Proof of Theorem 4.

The proof is omitted as it is almost identical to that of Theorem 1 in the special case with $\Psi(x,\mu) = \Phi(x-\mu)$.

References

Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference points and effort provision. American Economic Review 101, 470–492.

Beggs, A., 2005. On the convergence of reinforcement learning. Journal of Economic Theory 122, 1–36.

Benveniste, A., Metivier, M., Priouret, P., 1990. Adaptive Algorithms and Stochastic Approximation. Springer-Verlag, Berlin and New York.

Borkar, V., 2008. Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press, Cambridge.

Buhlmann, H., 1970. Mathematical Methods of Risk Theory. Springer Verlag, Berlin.

Camerer, C., Babcock, L., Loewenstein, G., Thaler, R., 1997. Labor supply of new city cabdrivers: One day at a time. Quarterly Journal of Economics 112.

Caplin, A., Dean, M., Glimcher, P., Rutledge, R., 2010. Measuring beliefs and rewards: A neuroeconomic approach. Quarterly Journal of Economics 125, 923–960.

Charpentier, A., Élie, R., Remlinger, C., 2021. Reinforcement learning in economics and finance. Computational Economics URL: `https://doi.org/10.1007/s10614-021-10119-4`.

Chew, S., 1989. Axiomatic theories of utility with the betweenness property. Annals of Operations Research 19, 273–298.

Cho, I.K., Rubinchik, A., 2017. Contemplation vs. intuition: a reinforcement learning perspective. EURO Journal on Decision Processes 5, 141–167.

Clarke, F., 1990. Optimization and Nonsmooth Analysis. Classics in Applied Mathematics, SIAM, Philadelphia, PA.

Dayan, P., Abbott, L., 2001. Theoretical Neuroscience. MIT Press, Cambridge, MA.

Dekel, E., 1986. An axiomatic characterization of preferences under uncertainty: relaxing the independence axiom. Jiournal of Economic Theory 40, 304–318.

Epstein, L., Zin, S., 1989. Substitution, risk aversion, and the temporal behavior of asset returns: A theoretical framework. Econometrica 57, 937–69.

Epstein, L., Zin, S., 1990. 'first-order risk aversion' and the equity premium puzzle. Journal of Financial Economics 26, 387–407.

Erev, I., Roth, A., 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review 88, 848–881.

Gill, D., Prowse, V., 2012. A structural analysis of disappointment aversion in a model of real effort competition. American Economic Review 102, 468–503.

Glimcher, P., 2011. Foundations of Neuroeconomic Analysis. Oxford University Press, Oxford.

Gosavi, A., 2006. Boundedness of iterates in q-learning. Systems and Control Letters 55, 347–349.

Gul, F., 1991. A theory of disappointment aversion. Econometrica 59, 667–686.

Hansen, L., Sargent, T., 1995. Discounted linear exponential quadratic gaussian control. IEEE Transactions on Automatic Control 40, 968–71.

Hansen, L., Sargent, T., 2008. Robustness. Princeton University Press, Princeton, NJ.

Hill, E., Bardoscia, M., Turrell, A., 2021. Solving heterogeneous general equilibrium economic models with deep reinforcement learning. URL: `https://arxiv.org/abs/2103.16977v1`.

Hopkins, E., Posch, M., 2005. Attainability of boundary points under reinforcement learning. Games and Economic Behavior 53, 110–125.

Johansen, T., Donaldson, J., 1985. The structure of intertemporal preferences under uncertainty and time consistent plans. Econometrica 53, 1451–1458.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 263–292.

Kimball, M., 1990. Precautionary saving in the small and the large. Econometrica 58, 53–73.

Köszegi, B., Rabin, M., 2006. A model of reference-dependent preferences. Quarterly Journal of Economics 121, 1133–1165.

Köszegi, B., Rabin, M., 2007. Reference-dependent risk attitudes. American Economic Review 97, 1047–1073.

Köszegi, B., Rabin, M., 2009. Reference-dependent consumption plans. American Economic Review 99, 909–936.

Kreps, D., Porteus, E., 1978. Temporal resolution of uncertainty and dynamic choice theory. Econometrica 46, 185–200.

Kuhnen, C., 2015. Asymmetric learning from financial information. Journal of Finance LXX, 2029–2062.

Kushner, H., Yin, G., 1997. Stochastic Approximation Algorithms and Applications. Springer Verlag, New York.

LaValle, I., 1968. On cash equivalents and information evaluation under uncertainty: Part i: Basic theory. Journal of the American Statistical Association 63, 253–276.

Leslie, D., Collins, E., 2005. Individual Q-learning in normal form games. SIAM Journal on Control and Optimization 44, 495–514.

Ma, D.J., Makowski, A., Shwartz, A., 1990. Stochastic approximations for finite-state markov chains. Stochastic Processes and their Applications 35, 27–45.

Marinacci, M., Montrucchio, L., 2010. Unique solutions for stochastic recursive utility. Journal of Economic Theory 145, 1776–1804.

Masatlioglu, Y., Raymond, C., 2014. A Behavioral Analysis of Stochastic Reference Dependence. Technical Report. University of Michigan.

Mihatsch, O., Neuneier, R., 2002. Risk-sensitive reinforcement learning. Machine Learning 49, 267–90.

Newey, W., Powell, J., 1987. Asymmetric least squares estimation and testing. Econometrica 55, 819–847.

Niv, Y., Edlund, J., Dayan, P., O'Doherty, J., 2012. Neural prediction errors reveal a risk-sensitive learning process in the human brain. Journal of Neuroscience 32, 551–562.

Niv, Y., Montague, R., 2009. Theoretical and empirical studies of learning, in: Glimcher, P., Camerer, C., Fehr, C., Poldrack, R. (Eds.), Neuroeconomics: Decision Making and the Brain. Elsevier, Amsterdam. chapter 22, pp. 331–351.

Perkins, S., Leslie, D., 2012. Asynchronous stochastic approximation with differential inclusions. Stochastic Systems 2, 409–446.

Pope, R., Chavas, J.P., 1985. Producer surplus and risk. Quarterly Journal of Economics 100, 853–869.

Pratt, J., 1964. Risk aversion in the small and in the large. Econometrica 32, 122–136.

Raiffa, H., 1968. Decision Analysis: introductory lectures on choices under uncertainty. Addison Wesley, Reading, MA.

Rouche, N., Habets, P., LaLoy, M., 1977. Stability Theory by Lyapunov's Direct Method. volume 22 of *Applied Mathematical Sciences*. Springer Verlag, New York.

Rusczyński, A., 2010. Risk-averse dynamic programming for markov decision processes. Mathematical Programming Series B 125, 235–261.

Sarver, T., 2012. Optimal Reference Points and Anticipation. Technical Report. Northwestern.

Schultz, W., 1998. Predictive reward signal of dopamine neurons. Journal of Neurophysiology 80, 1–27.

Shapiro, A., Dentcheva, D., Rusczyński, A., 2014. Lectures on Stochastic Programming: Modeling and Theory. Second ed., SIAM.

Shen, Y., Tobia, M., Sommer, T., Obermayer, K., 2014. Risk-sensitive reinforcement learning. ArXiv::1311.2097.

Singh, A., Jaakkola, T., Littman, M., Szepesvári, C., 2000. Convergence results for single step on-policy reinforcement learning algorithms. Machine Learning 39, 287–308.

Stokey, N., Lucas, R., Prescott, E., 1989. Recursive Methods in Economic Dynamics. Harvard University Press, Cambridge, MA.

Sutton, R., Barto, A., 2020. Reinforcement Learning. Second ed., MIT Press, Cambridge, MA.

Whittle, P., 1996. Optimal Control: Basics and Beyond. Wiley, New York.

Yaari, M., 1987. A dual theory of choice under risk. Econometica 55, 95–115.